Understanding Health Information Intent via Crowdsourcing: Challenges and Opportunities

Di Lu, University of Pittsburgh
Yihan Lu, University of Pittsburgh
Wei Jeng, University of Pittsburgh
Rosta Farzan, University of Pittsburgh
Yu-Ru Lin, University of Pittsburgh

Abstract
Social Q&A sites have been emerging as a platform for people to seek information and social supports around health topics. Identifying users' information needs from the questions can significantly help social Q&A sites serve their users better. Prior research had attempted to understand askers' intentions and implicit needs by classifying hidden intent from questions, while the non-trivial categorization was only able to be conducted with a limited size of data. In this study, we aim to develop a scalable categorization method that can categorize the askers' intent in a large set of health-related questions via crowdsourcing. We conducted a preliminary experiment on Amazon Mechanical Turk to evaluate our categorization method. Our results suggests both challenges and opportunities for understanding health information intent via crowdsourcing.
Keywords: social Q&A; crowdsourcing; categorization; health information behavior; content analysis
Citation: Lu, D., Lu, Y., Jeng, W., Farzan, R., Lin, Y-R. (2015). Understanding Health Information Intent via Crowdsourcing: Challenges and Opportunities. In *iConference 2015 Proceedings*.

Contact: dil16@pitt.edu, yil126@pitt.edu, wej9@pitt.edu, rfarzan@pitt.edu, yurulin@pitt.edu

1 Introduction

Online social Q&A sites, such as Yahoo! Answers, have become a prevailing cyberspace for people to seek health information or related supports (S. Oh, 2012). Unlike traditional web search engines, social Q&A websites provide more opportunities for users to express complex questions or needs by including more personalized or contextual information beyond simple queries. This uniqueness of social Q&A websites allows users to not only seek basic health information, but also address personal health issues or seek health-related social support. While prior research has developed a taxonomy of questions focusing on a specific health topic based on askers' intentions and implicit needs (Bowler, Oh, He, Mattern, & Jeng, 2012), classifying the hidden intent from questions is a non-trivial task – substantial training is required even for human coders to understand how to code askers' intentions. This costly process makes the categorization only possible to be applied to a limited size of data. However, the ability to reliably categorize questions can significantly improve how Q&A sites support their users. Thus, developing a simpler and still reliable categorization framework for health-related questions that can be easily adopted in large datasets becomes an important issue.

Crowdsourcing, on the other hand, defined as the process of obtaining needed service, ideas or content from the contribution of a large group of people, has been adopted by many research (Estellés-Arolas & González-Ladrón-de Guevara, 2012). Crowdsourcing service platform such as Amazon Mechanical Turk has been widely used to collect responses of simple tasks such as labeling pictures or calculating subtotal of bills (Kittur, Chi, & Suh, 2008). However, the potential opportunities of using crowdsourcing techniques to accomplish non-simple categorization tasks are still unclear. In this project, we aim to develop an effective method for understanding and categorizing the asker's intent in a large set of health-related questions via crowdsourcing.

This paper presents our study on developing a *scalable* categorization method that makes the task – categorizing health-related information seeking – scale out on the crowdsourcing platform. Our study can shed lights on the design of crowdsourcing tasks for non-trivial content analysis and for developing more sophisticated machine learning techniques. We discuss the challenges and opportunities of our study.

2 Method

2.1 Data Collection

Utilizing the Yahoo! Query Language (YQL) in Yahoo API[1], we collected the questions uploaded on Yahoo! Answers website under the Health category from May 1, 2014 to June 1, 2014 through an iterative crawling process. Overall, we created a Q&A corpus including 243,889 questions and 1,705,426 answers. The questions came from 4,723 users who we have their complete posting history until July 1, 2014. The percentage of questions in different categories are shown in Table 1.

Category	Question percentage
Entertainment & Music	22.44%
Society & Culture	8.60%
Pregnancy & Parenting	8.14%
Health	7.69%
Family & Relationships	7.21%
Beauty & Style	4.90%
Politics & Government	4.78%
Computers & Internet	4.16%
Arts & Humanities	3.44%
Science & Mathematics	3.24%

Table 1: Percentage of the questions in the top-10 categories.

2.2 Study Design

2.2.1 Coding scheme derived from prior work

As identified by prior research, informational and conversational questions are two main types of questions with different motivations and needs in online social Q&A websites (Harper, Moy, & Konstan, 2009). In particular, health-related questions in online social Q&A websites can be also classified into two similar broad types: informational and socio-emotional (J. S. Oh, He, Jeng, Mattern, & Bowler, 2013), based on the need and motivations of askers. Although both Bowler et al. (2012) and J. S. Oh et al. (2013) focused on a specific health topic – "eating disorder" and our work studies health-related questions in general, we believe the latent information needs in health-related questions are likely to be close. Thus, we decided to derived our categorization scheme from their work.

2.2.2 Open coding process, pilot coding and code book

We started with an open coding process in our study to evaluate the compatibility of the prior coding scheme in categorizing general health-related questions. Two researchers labeled every single questions regarding the intentions of the askers for 10 randomly selected questions from our entire dataset. The results of open coding process suggested that the coding scheme in prior study focusing on one certain health topic is compilable with categorizing general health-related questions.

Next, to make the coding scheme feasibly utilized through crowdsourcing, we modified the binary coding scheme into four categories regarding the primary intention of the askers: (1) Information seeking, (2) Social-support seeking[2], (3) Equally present, and (4) None or irrelevant. We added two more categories since some questions in our dataset contains dual intentions which are hard to be classified into one single category in the binary coding scheme. In some circumstances, questions may also be irrelevant to health

[1] https://developer.yahoo.com/yql/

[2] We used the term "social-support seeking" to emphasize that the primary intention of askers in this category is to seek social supports from others in online Q&A communities. The term "socio-emotional" used by literature focused more on the linguistic features of this type of questions.

topics or contains neither information seeking nor social-support seeking intention, in which cases they need the category "None or irrelevant" to hold. The selection criteria of Information seeking and Social-support seeking questions are described based on the micro-level intentions extracted from our open coding results.

We conducted a pilot study in order to test whether our coding scheme is clear and sufficient to deliver reliable outcomes. We recruited eight coders and conducted four sessions of pilot coding with 60 randomly selected questions from our dataset. Each coder coded 15 questions and each question is coded by at least two coders. Based on the feedback of pilots, we elaborated our selection criteria and eventually generated the code book for coders. Table 2 illustrates the coding scheme and selection criteria in our code book with examples of health-related questions for each category.

Category	**Selection Criteria**	**Example**
Information seeking	The primary intention of the asker is to ask for help in the following aspects: seeking treatment, listing out symptoms for a diagnose, asking medical advice, nutrition tips, information about therapy, medicine instruction.	I keep grinding my teeth? is this caused by stress? it's been like half a year since i last grinded my teeth and that time esta no bueno por mi. just wondering
Social-support seeking	The primary intention of the asker is to ask for comforts and sympathy, seek for an emotional outlet, seek for approval, express one's self, seek for a chat.	How can i make myself feel better? I've been feeling crap and lonely for a long time. I mean i just can't be bothered to keep in contact with people because none of them ever bother to get in touch with me. I've just lost hope. I just feel as though No-one can be bothered with me and no-one gives a sh8t. No matter how much i text get in touch with people no-one will ever like me enough to hang around or even get in touch with me without me having to call 1st. I spend a lot of time alone. To top that off i've never been in a proper relationship before and i'm always poor. Everyone i meet just ends up hurting me and doesn't want to know me. I don't know what to do with my life anymore. It makes me wonder whether i'll be lonely for the rest of my life and wonder why do i bother even living when life will always be sh8t.
Equally present	The question contains dual intentions as described in above two categories and the primary intention of the asker in hard to determine.	Please help me!! Im TERRIFIED of.....? WASPS!!! BEES!! HORNETS!!!! Anything that stings me! Please help, I can barely go outside without running right back inside like a little girl!!!!! Hahaha Everybody says dont bother it its more scared of you than you are of it if you leave it alone it will leave you alone But I DONT BELEIVE them!! So please help me
None or irrelevant	The question contains none of the intentions described in above categories or cannot be considered as a question related to health topics.	yay!!!!!!!!!!? someone messed up the computer systems at work and now I get to stay for another 3 hours! why am I not upset about this? i work at a hospital, lots and lots of people. anyone couldda done it...

Table 2: Coding scheme for categorizing different type of health-related questions and examples.

2.2.3 Preliminary experiment in Amazon Mechanical Turk

In order to test the feasibility of categorizing the information intent of health-related information through crowdsourcing, we conducted a preliminary categorization experiment on Amazon Mechanical Turk (AMT). AMT is one of the most popular crowdsourcing Internet marketplace that coordinates the use of human intelligence to perform tasks. On AMT, *Requesters* can post HITs (Human Intelligence Tasks) on the marketplace and *Workers* can then browse among existing tasks and complete them for a monetary payment set by the Requester. We launched three HITs in this experiment, each consists of 15 health-related questions and requests three independent turkers to categorize the questions based on the asker's intention according to the code book we provided. The duration limit for each Worker to complete each HITs is 15 minutes. To ensure the Worker's understanding of our tasks and their ability to complete the tasks, we created a qualification test that required Workers to categorize 10 questions before they can apply for our HITs. The threshold of accuracy for Workers to pass the qualification test is .80, that Workers need to categorize 8 out of 10 questions correctly to be qualified to work on our HITs. Moreover, in order to guarantee the quality of qualified Workers' responses of our HITs, we set two quality control questions hidden in each HITs. Workers need to categorize at least one of the quality control questions correctly in each HIT in order to get their responses approved. The "ground truth" of both the questions in qualification test and quality control questions in HITs are derived from an independent coding of two researchers (inter-coder agreement is .90).

3 Results and Discussion

Within two days after launching our HITs on Amazon Mechanical Turk, we received responses from nine individual Workers, three for each HIT. The inter-rater reliability (Krippendorff's alpha) among the three coders for each of the three HITs (.4044, .4821, .4159) are lower than the minimum acceptable reliability in social sciences (.667)(Krippendorff, 2012).

We found that the most disagreements among coders were caused by the different perception of the boundary of category "Equally present". Some coders tended to categorize more questions into category "Equally present" since they failed to identify the primary intention if both types of intention existed in a single question, whereas other coders may be able to classify a dual-intentional question into either "Information seeking" or "Social-support seeking" by identifying the primary intention of the asker. Thus, we decided to modify our coding scheme by refining the boundaries between different categories. We plan to separate original category "Equally present" into two subcategories based on the primary intention in a dual-intentional question. In this way, coders have to identify the primary intention when they found both information seeking and social-support seeking intentions in a question. The category "Information seeking" and "Social-support seeking" are revised as referring to questions contains purely one type of intentions. More specifically, we received more results that were categorized as information seeking rather than social-support seeking.

Generally, our preliminary experiment on AMT received fast responses from Workers with low expenses. Our results suggest valuable opportunities for understanding the information seeking in diverse contexts from a large-scale content analysis via crowdsourcing and make it possible for machine learning techniques. However, our study also identified several critical challenges for developing categorization framework for non-trivial categorization tasks via crowdsourching. First, an important challenge for large-scale content analysis is to control the quality of the responses. Although we adopted the qualification test to ensure Worker's ability to accomplish the tasks and hided quality control questions in each HIT to evaluate the reliability of Workers' responses, the inner-rater reliability of every three coders of each HIT in our experiment is still below the minimum acceptable threshold. The cause of the low reliability introduced another challenge – how to draw clear boundaries between categories for non-trivial categorization tasks involving overlapping concept so that different coders can perceive it consistently. Further, the consistent perception of the boundaries between categories is also important for training classifiers with machine learning techniques in the future. Besides, the skew distribution of askers' intentions in our dataset makes the quality control and measurement of reliability more challenging. Since social Q&A sites, by nature, provide a space for users to express information needs (Shah, Oh, & Oh, 2009), the askers' intention to seek information is presented in most of the questions in our dataset. The dominance of one category raises potential bias which may causes the results of inter-rater reliability tests to misrepresent the inter-rater reliability among coders since it is hard to judge whether the tendency for coders to identify one category more than others is due to

the skew distribution in our dataset or the poor performance of coders. In future work, we plan to design AMT experiments that take these considerations into account, for example, by adding more sophisticated qualification test.

References

Bowler, L., Oh, J. S., He, D., Mattern, E., & Jeng, W. (2012). Eating disorder questions in yahoo! answers: Information, conversation, or reflection? *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1–11.

Estellés-Arolas, E., & González-Ladrón-de Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information science*, *38*(2), 189–200.

Harper, F. M., Moy, D., & Konstan, J. A. (2009). Facts or friends?: distinguishing informational and conversational questions in social q&a sites. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 759–768).

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with mechanical turk. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 453–456).

Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*. Sage.

Oh, J. S., He, D., Jeng, W., Mattern, E., & Bowler, L. (2013). Linguistic characteristics of eating disorder questions on yahoo! answers–content, style, and emotion. *Proceedings of the American Society for Information Science and Technology*, *50*(1), 1–10.

Oh, S. (2012). The characteristics and motivations of health answerers for sharing information, knowledge, and experiences in online environments. *Journal of the American Society for Information Science and Technology*, *63*(3), 543–557.

Shah, C., Oh, S., & Oh, J. S. (2009). Research agenda for social q&a. *Library & Information Science Research*, *31*(4), 205–209.

Table of Figures

Table of Tables

Table 1 Percentage of the questions in the top-10 categories. 2
Table 2 Coding scheme for categorizing different type of health-related questions and examples. 3